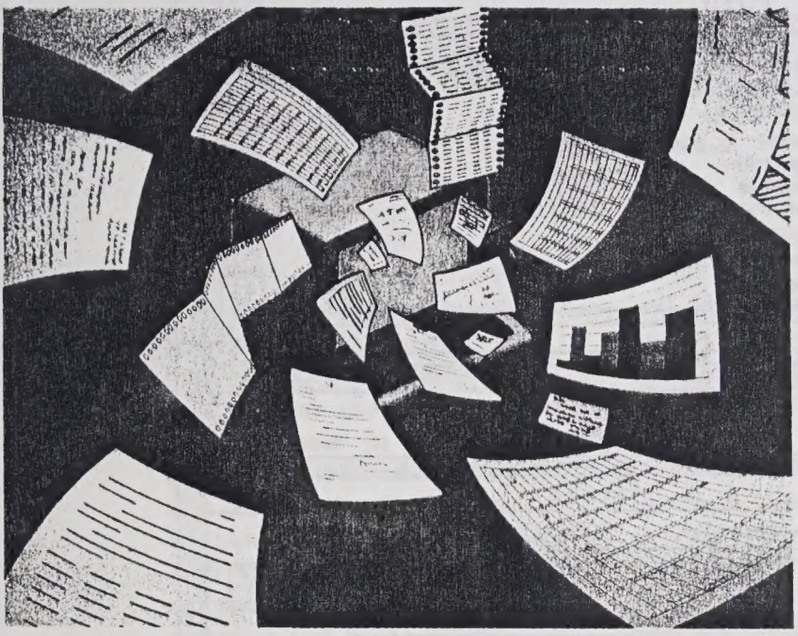
Collective Dynabases



ILLUSTRATIONS: BARTON STABLER

In my last column [15], I discussed portable computers as they are used for input to a person's dynabase, a dynamic database containing the notes, sketches, papers, and other documents he or she creates and gathers over a long period of time. The focus was on an individual's dynabase, but what about collective dynabases? For example, could the dynabases of everyone who worked for a company be combined, producing a corporate dynabase?

Much of the information an organization acquires or produces is already in machine-readable form. For example, text documents such as memos, letters, and reports are prepared on computers, and networks capture electronic mail. As portable computers and networks improve and proliferate, the percentage of information in machine-readable form will increase, making it still easier to create a dynabase as a byproduct of everyday work.[1]

Once information is in the dynabase, however, how will it be viewed and retrieved? To what extent will users have to add retrieval-enhancing information such as key words? To what extent will the system be able to generate such clues by analyzing a document or the context in which it was created? These are difficult questions, but if they sound insurmountable, do not despair—the system does not have to be perfect, just better than what we have today.

Researchers and commercial computing people are working on the challenge of collective dynabases. Let us look at examples of both, beginning with a research project on wide area information servers (WAIS).

Wide Area Information Servers

The WAIS project is headed by Brewster Kahle at Thinking Machines. Thinking Machines makes Connection Machines—highly parallel supercomputers that are well suited to free-text search [19]. In most text-retrieval systems, queries are limited to Boolean combinations of a few terms, but since text on a Connection Machine is fast, searches for documents which are similar to an entire document are practical. This technique is used in Dow Jones's DowQuest, a commercial system that uses a Connection Machine to scan more than 150,000 articles from 185 publications for relevance to online queries. While DowQuest is proprietary, the WAIS project is an attempt to open the technology.

WAIS is a client-server system, and as of this writing, there are 193 WAIS servers on the Internet, covering topics from poetry to television programs. Past issues of *Communications* are now available on a server. There is public domain client software for the Mac, DOS, Windows, NeXT and others. Since a standard protocol has been devised for queries and replies, any client can be used with any server. The protocol extends the NISO Z39.50 protocol developed for library catalog queries, and it is being used and developed by an active group of Internet-based researchers [12].[2]

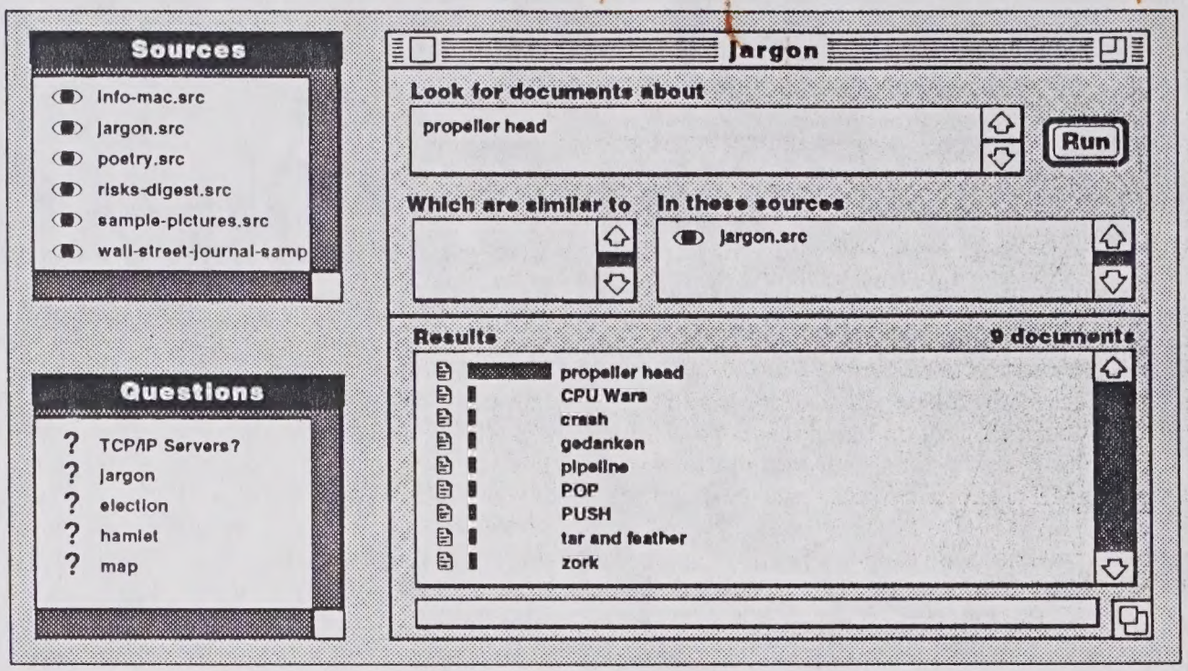
A WAIS session typically proceeds in two phases. First, you do a standard keyword search, identifying several documents. Then you search for documents similar to the most relevant ones from the first phase. Figure 1 shows a WAIS session using a Mac client. The session began with a query that sought documents about "propeller head" on the jargon.src server (Figure 1a). (Jargon.src contains the definitions found in [17], a delightful compendium of hacker jargon.) Nine documents were found, and their titles listed. Note that the length of the gray bar before the listed titles indicates the strength

Larry Press

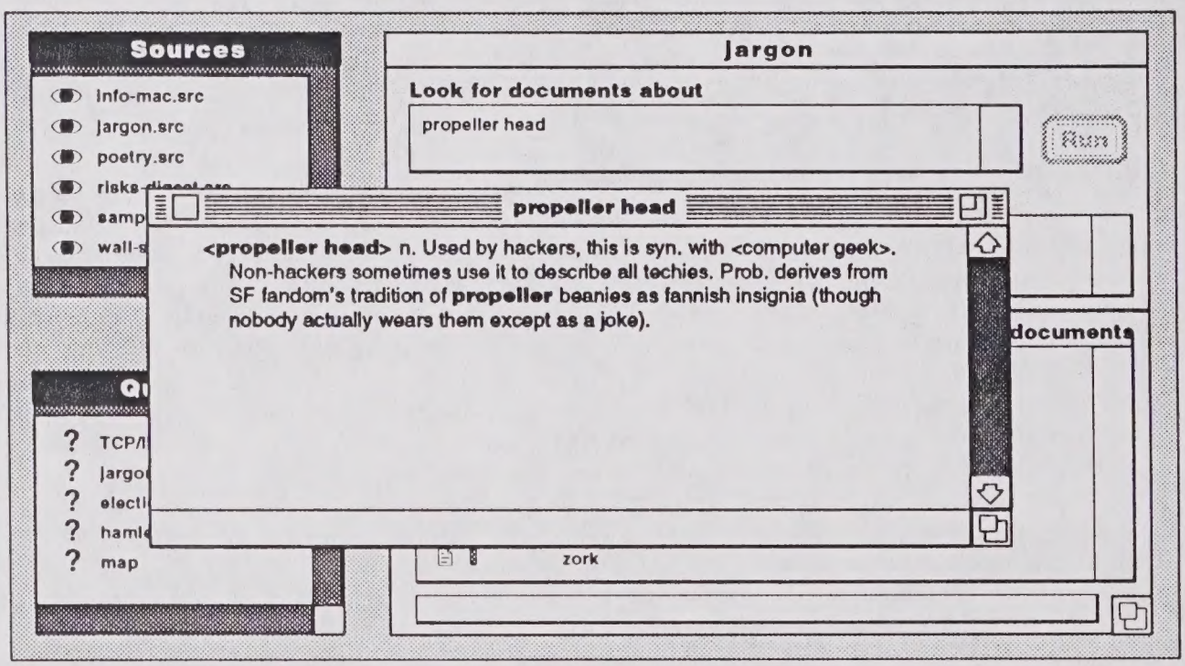
[1] Even if no extra effort is needed to enter information, individuals would have to give permission for information they create or gather to be published in the collective dynabase. Other issues affecting an individual's willingness to participate in such systems are discussed in [7, 15].

[2] Two other protocols, SFQL and CD-RDx, are also being developed for decoupling database queries. Structured Full-Text Query Language (SFQL) extends SQL, and it grew out of the need to query heterogeneous aircraft documentation distributed on CD-ROM [18]. It is being developed at the request of the Air Transport Association. CD-RDx is being developed at the request of the Information Handling Committee of the CIA (see [22]). Their goal is to enable government agencies to share data.

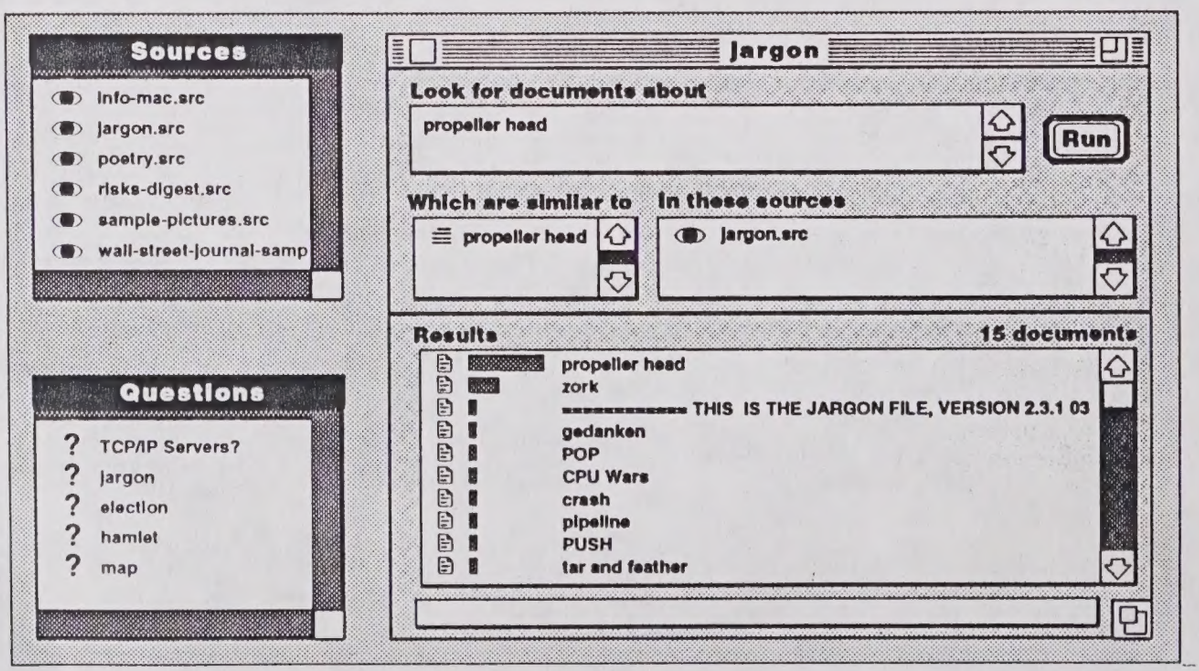
Figure 1. A WAIS session. This session involves two queries. First a keyword search for "propeller head" is run on the jargon.src database (1a). Nine documents are found, including one entitled "propeller head" (1b). The second search retrieves documents that are *similar to* the one called "propeller head" (1c).


Sources
Info-mac.src
Jargon.src
poetry.src
risks-digest.src
sample-pictures.src
wall-street-journal-samp
Questions
TCP/IP Servers?
Jargon
election
hamlet
map
Jargon
Look for documents about
propeller head
Run
Which are similar to
In these sources
Jargon.src
Results
9 documents
propeller head
CPU Wars
crash
gedanken
pipeline
POP
PUSH
tar and feather
zork

A


Sources
Info-mac.src
Jargon.src
poetry.src
Jargon
Look for documents about
propeller head
Run
propeller head
<propeller head> n. Used by hackers, this is syn. with <computer geek>. Non-hackers sometimes use it to describe all techies. Prob. derives from SF fandom's tradition of propeller beanies as fannish insignia (though nobody actually wears them except as a joke).
zork
map

B


Sources
Info-mac.src
Jargon.src
poetry.src
risks-digest.src
sample-pictures.src
wall-street-journal-samp
Questions
TCP/IP Servers?
Jargon
election
hamlet
map
Jargon
Look for documents about
propeller head
Run
Which are similar to
propeller head
In these sources
Jargon.src
Results
15 documents
propeller head
zork
THIS IS THE JARGON FILE, VERSION 2.3.1 03
gedanken
POP
CPU Wars
crash
pipeline
PUSH
tar and feather

C

of a heuristic index of the degree of relevance. Any of the documents can be retrieved by clicking on it, and Figure 1b shows the document entitled "propeller head." Figure 1c shows the result of a second search, in which I asked it to find documents *similar to* the document called "propeller head." As you see, it found 15 such documents.

Although WAIS is experimental, it is easy to use, and the fact that the servers are spread over the Internet is well hidden. Automatic searches (e.g., a daily search of the *Wall Street Journal*) are also possible and bit maps and sound files are supported.

WAIS was also used as a corporate dynabase in a three-month experiment, run with the cooperation of Apple, Dow Jones, and Thinking Machines at the KPMG Peat Marwick consulting firm [9]. According to Robin Palmer, senior manager and WAIS project leader, WAIS integrated internal information such as word processing documents, letters, reports to clients, and training materials with external data like DowQuest and technical tax and accounting literature from the Financial Accounting Standards Board. Palmer is quite positive. In one instance they prepared a proposal for a new prospect on four hours notice. It included competitive and industry analysis and a discussion of related work by KPMG, and led to a contract on the first meeting. Palmer felt the internal data was most important in that it "reflected KPMG's internal knowledge." He warns, however, that three weeks were needed for initial data conversion, and that WAIS requires at least 56Kb/sec communication.

The Internet experiments and KPMG test are research prototypes, but they foreshadow what may be possible when networks are ubiquitous and Connection Machines are on desktops. Let us now look at some commercial developments.

Commercial Beginnings—Lotus Notes

Lotus is best known for its 1-2-3 spreadsheet, but many years ago it made a strategic commitment to groupware, and in 1989, began selling a system called Notes.[3] Notes is a step toward the corporate dynabase.

A company with only one location would see Notes as a LAN-based conferencing system with a graphical user interface (see Figure 2). The LAN would have a dedicated Notes server with many databases.[4] Notes databases contain documents of one or many formats, and field data types include dates, numbers, free text, and rich text (text with varying fonts and bit-mapped graphics and check-box categories).

Users at client stations post documents to databases by filling in screen forms, and if they have permission, view and edit them. The user interface conforms to IBM's Common User Access Standard, and users can search and view a database in many ways, for example selecting only certain records and fields to be displayed or changing their order. The database designer would provide some default views, and users are free to create custom views. Totals, averages and other statistics can be computed, and more complex operations may be programmed using the Notes API, which gives access to databases.

If this were all it did, Notes would be quite useful, but there is another key aspect. Notes is not restricted to single locations. It was designed for distributed organizations that do not have a persistent data link between sites [10]. If an organization had two or more locations, each would have a LAN with a Notes server. At time intervals set by the system administrator, connections would be established between sites, and the servers would exchange all new data, leaving the databases at both sites identical. Notes also supports users with portable or stand-alone computers who dial into a server. This database replication is similar to periodic updates between network directory servers, and while it would not work for transaction processing and other applications requiring immediate synchronization, it is sufficient for a corporate dynabase.

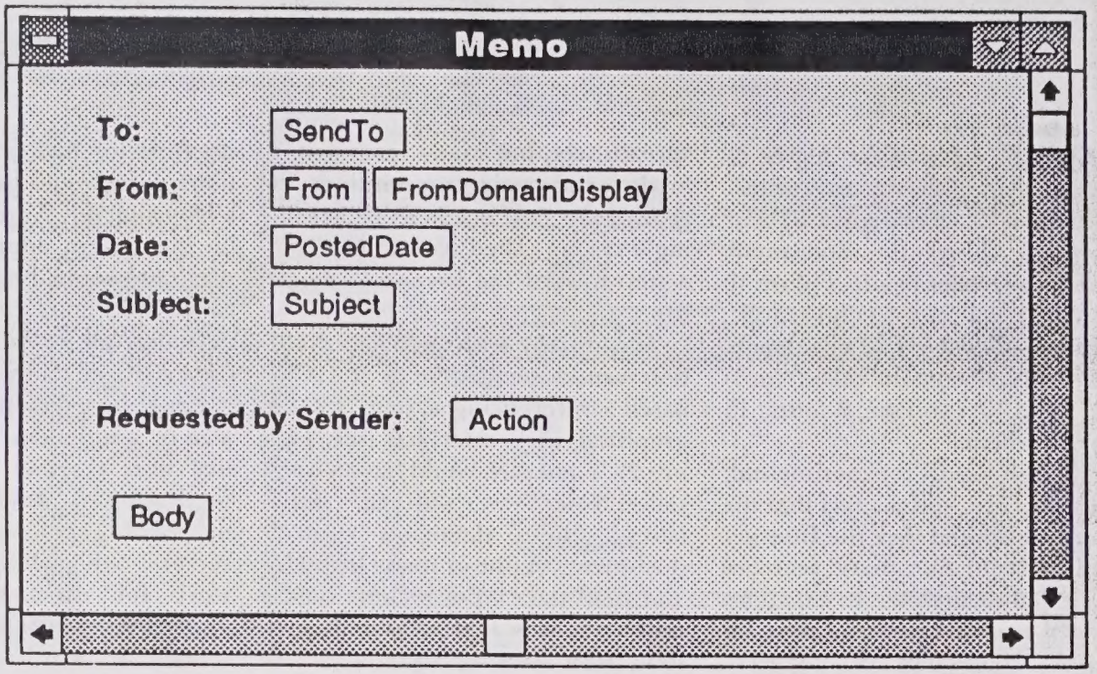
Notes is sold to large organizations (the minimum installation is 200 users at a price of $62,500), and there is now considerable experience with it. For example, Marshak reports that Notes is being used to create an environment of information and processes at Manufacturers Hanover Trust, a large New York bank [13]. The bank's goal is to "get the information stored in people's heads into a form and place where others can use it," and their applications include a database of customer and potential customer profiles and a database of communications with customers. Marshak also reports on the Price Waterhouse accounting firm. Of its 12,000 employees, 6,000 are Notes

[3]It is interesting to speculate on how companies make such decisions. Perhaps proximity to a nearby research project plays a role. Lotus was in Cambridge, and hired groupware researcher Irene Greif from MIT, and Apple was near Xerox PARC, where they got ideas and people for the Lisa and Mac. Notes was developed for Lotus by Ray Ozzie who was hired for another product, but had wanted to build a conferencing system since his college days. How did MicroSoft decide to move into CD-ROM and multimedia?

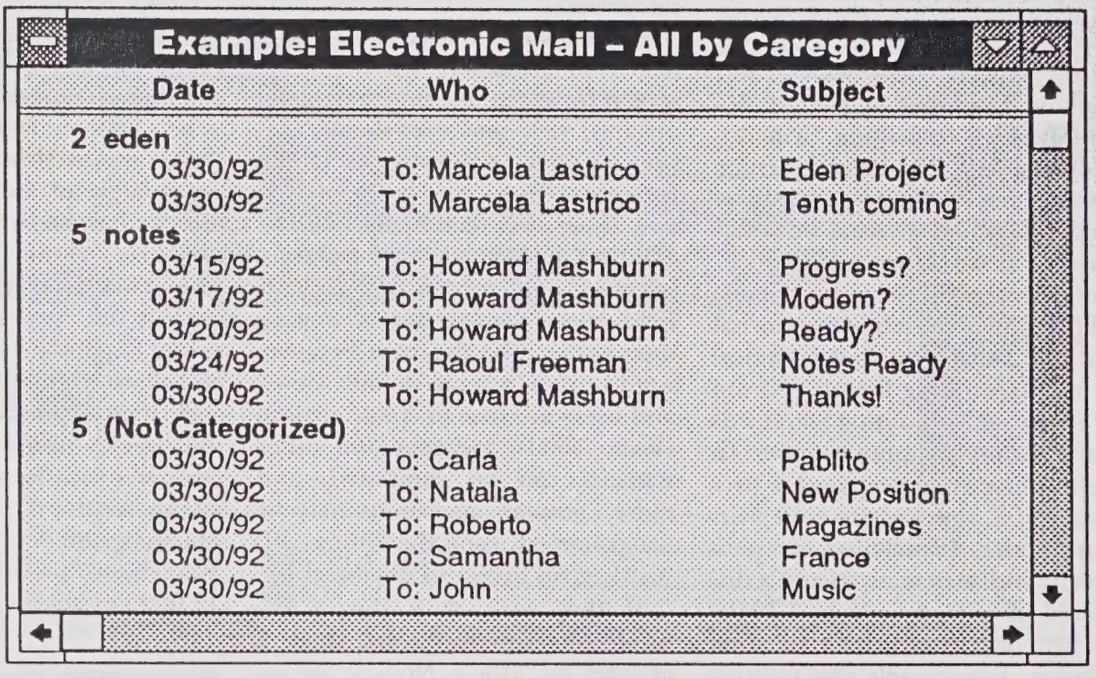
[4]Today Lotus has only OS/2-based servers and Windows and OS/2-based clients, but servers for Unix, Netware and NT and Mac and Unix clients are under development.

Figure 2. Lotus Notes. This database holds electronic mail messages sent and received. When a database is defined, one or more forms are created (2a). For example, a mail database might have records for general messages (shown here), replies, and phone messages. The designer specifies form layout and access control, field data types, default values, formulas for validity checking and optional value computation, and other field and form properties. Views are also defined. For example, (2b) shows messages sorted by user-assigned category, but others like view by name or data could also be defined. In 2c, the user has displayed one record. Note that date and category check-boxes are built-in types, and the rich-text field with the message body contains a bit-mapped drawing.

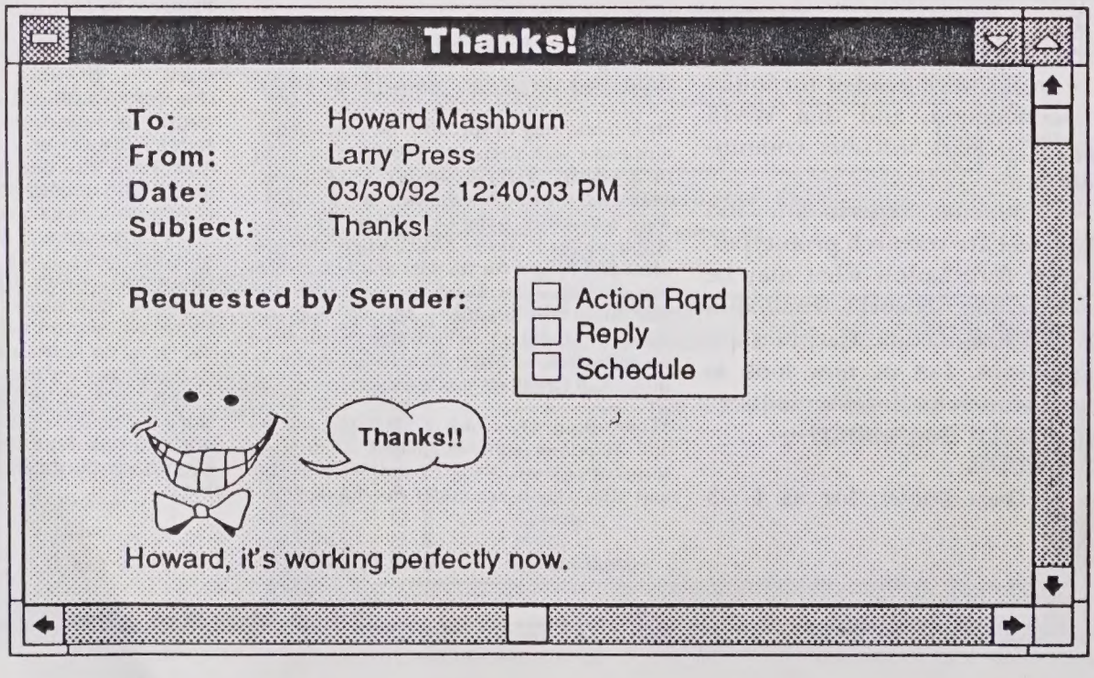
A


Memo
To:
SendTo
From:
From
FromDomainDisplay
Date:
PostedDate
Subject:
Subject
Requested by Sender:
Action
Body

B


Example: Electronic Mail – All by Caregory
Date
Who
Subject
2 eden
03/30/92
To: Marcela Lastrico
Eden Project
03/30/92
To: Marcela Lastrico
Tenth coming
5 notes
03/15/92
To: Howard Mashburn
Progress?
03/17/92
To: Howard Mashburn
Modem?
03/20/92
To: Howard Mashburn
Ready?
03/24/92
To: Raoul Freeman
Notes Ready
03/30/92
To: Howard Mashburn
Thanks!
5 (Not Categorized)
03/30/92
To: Carla
Pablito
03/30/92
To: Natalia
New Position
03/30/92
To: Roberto
Magazines
03/30/92
To: Samantha
France
03/30/92
To: John
Music

C


Thanks!
To:
Howard Mashburn
From:
Larry Press
Date:
03/30/92 12:40:03 PM
Subject:
Thanks!
Requested by Sender:
Action Rqrd
Reply
Schedule
Thanks!!
Howard, it's working perfectly now.

users. While only one-third of their top management had used PCs before Notes was installed, all are now active Notes users. Notes is seen as a strategic tool, and the Price Waterhouse chairperson characterized one of the firm's major challenges as the need to collect, manage, and access "the expertise of our people."

We shape our tools and they shape us. Information processing tools and organizations co-evolve, and Notes affects and is affected by corporate culture. Ed McDonald, manager of the Information Processing Division at Texaco Oil, states that they have 1,500 Notes users. Before they began using Notes, Texaco had decided to alter the corporate shape and culture, moving from a rigid hierarchy to a flatter organization with freer lines of communication and more information sharing. McDonald feels Notes would not have been as well received without this move, and that, in turn, it has facilitated the shift. As an example, before Notes, his subordinates sent him weekly activity reports. Now they post the reports on Notes, and everyone can see what the others are doing (and how hard they are working).

Texaco also uses Notes for distributed synchronous meetings. Each month, 20 people from eight U.S. locations meet, using an audio link and Notes. Notes is used to distribute the agenda before the meeting and for note taking and brainstorming during the meeting (each user logs into the same server). The major Notes controversy is a debate over the use of company resources to support an informal "coffee bar" conference where people shoot the breeze as when on coffee breaks.

With time, we will gather systematic data on the impact of such systems. Brad Jackson of Texaco is collaborating with researchers at the University of Minnesota in a two-year study of the impact of group decision support tools including Notes. Wanda Orlikowski and her colleagues at MIT are studying the application and impact of Notes in several organizations. In the meantime, customers are installing Notes (e.g., Arthur Anderson has purchased 20,000 licenses and committed to 60,000), and it is a strategic product which Lotus is extending rapidly. Lotus has also formed key business alliances. For example, IBM markets Notes, Kodak is working with Lotus on document-imaging, Verity is adding their text-retrieval capability, and Notes is being implemented as a Novell Netware Loadable Module, which will eliminate the need for an OS/2 server (but may cut speed).

While Notes gives Lotus a lead, it is only a first step. People will not use a corporate dynabase if doing so is difficult. For a start, seams between the dynabase and other manual and software tools must be closed [8]. Lotus, MicroSoft and others are mail-enabling their applications. Spreadsheet and word processing users will soon be able to *mail* a document or portion of a document as easily as they now *print* or *save* it. Instead of importing files, users will be able to use their word processor while working in Notes. More integration is needed. A successful dynabase requires natural interfaces to OCR systems, portable computers, pens, faxs, telephones, and wide-area networks. Stitching seams will foster building the dynabase as a byproduct of other work.

Standards are also needed. Today Notes is self-contained, but it is being rewritten to run on top of VIM, Lotus's Vendor Independent Messaging interface. IBM, Apple, Novell, Borland, and others are committed to VIM. MicroSoft is not because they have their Messaging Application Program Interface (MAPI). Such standards will not only allow software vendors to mail-enable current applications, they will use them for new applications such as scheduling and workflow. In-house programmers will also use them for custom applications.

While Lotus is ahead with Notes, MicroSoft is ahead with MAPI. They have shipped a software developer's kit, and two-thirds of the over 300 people in their Workgroup Applications Business Unit are working on system software. MAPI and VIM overlap in the services they support, and it would be premature to guess which will (or should) prevail. Technical people from both camps are in contact. It took roughly 50 years to establish a standard, unified telephone system in the U.S. [14]. In the long run, communication services will be part of operating systems, and all applications will use them. As with the telephone, the standards will be decided by the market place and corporate power, not just technical merit.

While formidable, solutions to the problems of bridging seams and developing standards are understood. Dynabase retrieval and organizational problems are more open-ended. The ideal dynabase would have a simple interface. The user would present it with an arbitrary document, and either ask that it be filed away or that similar documents be retrieved. That vision calls to mind vague prospects for natural-language processing for classifying and indexing text or recognized speech documents and pattern recognition for graphics. While we are far from such capability, an interim dynabase can use heuristics and file inversion to index text (as WAIS does). It can relate items by the date and context

In the long run, communication services will be part of operating systems, and all applications will use them.

in which they were created (for example the words on a page surrounding a sketch or the project you were working on when you made it), and allow the user to assign key words, degrees of importance, and other explicit clues.

Beyond the Corporate Dynabase

Information systems cross corporate boundaries. For example, manufacturers send email to suppliers and customers, and Electronic Data Interchange (EDI) is used for intercompany transactions. The infrastructure being built to support such activity will be available for interorganizational dynabases.

Patricia Seybold's Office Computing Group has created an interorganization dynabase. Seybold has traditionally done consulting and published newsletters and research reports. Last year the group began publishing a Notes database containing the information in the printed versions of their four monthly newsletters plus special bulletins and analysis written for the Notes group. This information gets to clients immediately, and it can be searched and viewed in any way the reader desires. Note that this is not just electronic distribution of a newsletter (as if it were being faxed); it is a two-way link with clients who make comments and guide the direction of further research. Seybold has a sweeping vision, and expects most of her clients to eventually switch from printed to electronic reports within two to three years.

Ken Laws is not phasing out printed newsletters. His year-old "Computists Communique" was electronic from the start, and is available on the Internet.[5] Each issue covers news of the domestic and international computer and research industries with topics such as job opportunities, calls for papers, and conference announcements. Another section has pointers to newsletters, databases, software and other Computists' tools. These are followed by commentary and an in-depth essay on a related topic. There is something noteworthy (worth saving in your dynabase) in every issue. Similar to Seybold, Laws is creating a community. Issues are peppered with comments from and pointers to readers, and past issues are retrievable.

My last column discussed individual dynabases, fed from portable computers. This one has moved to organization and interorganization dynabases, but speculative writers consider the entire planet. The French paleontologist, theologian and philosopher Pierre Teilhard de Chardin [20] held that the matter-energy comprising the universe is constantly evolving in the direction of increased complexity, leading to the formation of the earth, the "geosphere," and life, the "biosphere." In 1925 he coined the term "noosphere"—an evolving network of human culture, connection, knowledge and interdependence. Zoologist Richard Dawkins [2] discusses "memes"—replicable ideas that live in people's minds. Dawkins holds that memes are not mere metaphors, but are literally realized in the physical structure of the nervous system of the individuals in which they reside. If the earth is an evolving organism, perhaps the Internet is a step in the evolution of the Gaian nervous system, and you and I and WAIS servers are neurons.

[5]Laws began electronic publishing as the moderator of the Internet AI list.

References:

1. Brin, D. *Earth.* Bantam Books, New York, 1990.
2. Dawkins, R. *The Selfish Gene.* Oxford University, Oxford, 1976.
3. Engelbart, D.C. Knowledge-domain interoperability and an open hyperdocument system. In *Proceedings of the ACM Conference on Computer-Supported Cooperative Work,* (Los Angeles, CA, Oct. 1990), 143–156.
4. Engelbart, D.C. and English, W.K. A research center for augmenting human intellect. In *Proceedings of the AFIPS Fall Joint Computer Conference* (San Francisco, Dec. 1968), pp. 395–410.
5. Erickson, T. and Salomon, G. Designing a desktop information system: Observations and issues. *CHI '91 Conference Proceedings,* Addison-Wesley, Reading, Mass., 1991, pp. 49–54.
6. Gorry, G.A., Long, K.B., Burger, A.M., Jung, C.P., and Meyer, B.D. The virtual notebook system™: An architecture for collaborative work. *J. Org. Comput. 1,* 3 (1991), 233–250.
7. Grudin, J. Why CSCW applications fail: Problems in the design and evaluations of organizational interfaces. In *Proceedings of the ACM Conference on Computer-Supported Cooperative Work* (Portland, Oreg., Sept. 1988).
8. Ishi, H. and Miyake, N. Toward an open, shared workspace. *Commun. ACM* (Dec. 1991), 37–50.
9. Kahle, B. and Medlar, A. An information system for corporate users: Wide area information servers. FTP: pub/wais/doc/wais-corp.txt@think.com, Apr. 8, 1991.
10. Kawell, L., Beckhardt, S., Halvorsen, T., Ozzie, R. and Greif, I. Replicated document management in a group communication system, in Marca D. and Boch, G. *Groupware: Software for Computer-Supported Cooperative Work.* IEEE Press, to be published.
11. Lovelock, J.E. *Gaia.* Oxford University, Oxford, 1979.
12. Lynch, C.A. The Z39.50 Information retrieval protocol: An overview and status report. *ACM SIGCOMM Comput. Commun. Rev. 21,* 1 (Jan. 1991), 58–70.
13. Marshak, D.S. Lotus notes, a platform for group information management applications. Special Research Report, Patricia Seybold's Office Computing Group, Boston, Mass., June, 1991.
14. Pierce, J.R. *Signals, The Telephone and Beyond.* Freeman, San Francisco, Calif., 1981.
15. Press, L. Dynabook revisited. *Commun. ACM* (Mar. 1992).
16. Press, L. Systems for finding people. *J. Org. Comput.,* to be published.
17. Raymond, E. *The New Hacker's Dictionary.* MIT Press, Cambridge, Mass., 1991.
18. Shapiro, N.R. Diamantopoulos, E.

and Cotton, P. CD-ROM Disc Interchangeability Standards: Beyond ISO 9660 with Structured Full-Text Query Language (SFQL), ATA/AIA 89-9C Monograph, April 1991.

19. Stanfill, C. and Kahle, B. Parallel Free-Text search on the Connection Machine system. *Commun. ACM 29*, 12 (Dec. 1986), 1229–1239.

20. Teilhard de Chardin, P. *The Phenomenon of Man.* Harper and Row, N.Y., 1955.

21. WAIS interface protocol, prototype functional specification. Apr. 1990. Anonymous FTP: /pub/wais/doc/protspec.txt@think.com.

22. Standard for the exchange of digital information on CD-ROM. DCI Information Handling Committee, Intelligence Community Staff, Wash., D.C., 20505, Version 3.1, May, 1991.

Since St. Jude Children's Research Hospital opened in 1962, it has forged new treatments for childhood cancer and has helped save the lives of thousands of children around the world. But the battle has just begun. You can join the fight. To find out how, call 1-800-877-5833.

ST. JUDE CHILDREN'S RESEARCH HOSPITAL
Danny Thomas, Founder

Pointers

• For more information on WAIS, contact info@think.com on the Internet. Papers about WAIS can be retrieved by anonymous ftp from /pub/wais/doc@think.com. To retrieve Communications of the ACM articles, log on and select the database called "cacm". If you do not have Internet access, Thinking Machines Corp. is at 245 First Street, Cambridge, MA 02142-1264; tel: (617) 234-1000; fax: (617) 234-4444. For a discussion of interface issues, WAIS, and DowQuest, see [4].

• The 1992 revision of the Hacker's Dictionary is 100% fun. It is the perfect programmer's aptitude test. If you feel like putting it down in less than one hour after picking it up, look for another profession. (This suggestion is only half joking —people with an aptitude for a symbol-manipulating profession such as programming might have similar taste in humor and wordplay).

• Patricia Seybold's Office Computing Group does consulting and publishes material about groupware, object-oriented technology, and a variety of other topics. This group covers technology, the industry, and impact on organizations. 148 State Street, 7th Floor, Boston, MA 02109; tel: (800) 826-2424; fax: (617) 742-1028; pseybold@mcimail.com.

• Science fiction fans who would like to pick up where this column ended, will enjoy David Brin's book *Earth 1*, 1990.

• Ken Laws, 4064 Sutherland Drive, Palo Alto, CA 94303; tel: (415) 493-7390; email: laws@ai.sri.com.

• Any discussion of corporate memory needs a pointer to Doug Engelbart. Engelbart inspired a generation of researchers when he demonstrated his NLS system at the 1968 Fall Joint Computer Conference. For a description of his early work see [4], and for his current views on interorganization dynabases, see [3]. Much of the 1968 demonstration is shown on the video tape of ACM's Conference on the History of Personal Workstations. Doug Engelbart, The Bootstrap Institute, 6505 Kaiser Drive, Fremont, CA 94555; tel: (510) 713-3550; fax: (510) 793-2362; email: engelbart@bootstrap.stanford.edu.

• Of course Notes and WAIS are not the only systems working toward organizational memory. Another noteworthy effort is the Virtual Notebook System™, a networked, multimedia system for recording lab notes and other research-related information [6]. Shared electronic notebooks contain text, image, audio or video entries, and notebook objects can be linked to other objects or to executable programs. The system is being developed at Baylor College of Medicine, and tested there and at several other sites. A commercial release is available from The Fore-Front Group, 1709 Dryden, Suite 901, Houston, TX 77030; tel: (713) 798-6116; fax: (713) 798-3729; email: klong@bcm.tmc.edu.

• Corporate and intercorporate dynabases are studied in schools of business. You might be interested in the annual Conference on Organizational Computing, Coordination and Collaboration, which brings together business school faculty and industrial practitioners in a workshop-like setting. The papers presented appear in special issues of the *Journal of Organizational Computing*, Ablex, Norwood, N.J. For information contact Andrew Whinston, abw@emx.utexas.edu.

Larry Press welcomes questions and comments from readers. His address, phone number and email are: 10726 Esther Avenue, Los Angeles, CA 90064; (310) 475-6515; lpress@venera.isi.edu

Larry Press is a professor of computer information systems at California State University at Dominguez Hills.